(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENT- UND MARKENAMT

(12) **Offenlegungsschrift**

(10) **DE 100 29 644 A 1**

(51) Int. Cl.[7]: G 06 F 17/30
G 06 F 3/00

(21) Aktenzeichen: 100 29 644.0
(22) Anmeldetag: 16. 6. 2000
(43) Offenlegungstag: 17. 1. 2002



DE 100 29 644 A 1

(71) Anmelder:
Deutsche Telekom AG, 53113 Bonn, DE

(72) Erfinder:
Hoppe, Thomas, Dr., 10585 Berlin, DE; Oertel, Helmut, 14055 Berlin, DE; Paulus, Oliver Kai, 10115 Berlin, DE; Ewert, Marc, 12099 Berlin, DE

(56) Für die Beurteilung der Patentfähigkeit in Betracht zu ziehende Druckschriften:

| | |
|---|---|
| DE | 198 42 320 A1 |
| DE | 197 29 911 A1 |
| DE | 196 51 788 A1 |
| US | 59 20 859 A |
| EP | 08 38 056 B1 |
| EP | 06 31 245 B1 |

**Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen**

(54) Verfahren zur Relevanzbewertung bei der Indexierung von Hypertext-Dokumenten mittels Suchmaschine

(57) Die Erfindung bezieht sich auf ein Verfahren zur Relevanzbewertung bei der Indexierung von Hypertext-Dokumenten mittels Suchmaschine, welches in drei Phasen abläuft. In der Aufbauphase liefert das Robotersystem Hypertext-Dokumente an den Indexserver. Der Indexserver analysiert den Inhalt der Dokumente nach drei unterschiedlichen Gesichtspunkten. In der Aktualisierungsphase werden Dokumente, deren Inhalte sich seit dem letzten Besuch verändert haben, zunächst aus dem Dokumenten-Index entfernt. Die betreffenden Termeinträge werden aktualisiert. Sofern das veränderte Dokument weiterhin verfügbar ist, wird entsprechend den Arbeitsschritten der Aufbauphase in den Index eingefügt. In der Anfragephase werden in Abhängigkeit vom verwendeten Anfragetyp (einfache Anfrage, komplexe Anfrage, Boolsche Anfrage oder Phrasenanfrage) aus dem Index die Dokumente ermittelt, die auf die Anfrage zutreffen. Für jedes gefundene Dokument wird der eigentliche Relevanzwert aus den vorausberechneten Relevanzwertanteilen, der zum Anfragezeitpunkt vorliegenden Anzahl an Verweisen auf das Dokument und der Gesamtanzahl der Dokumente im Index zum Relevanzwert des Dokuments verrechnet.

DE 100 29 644 A 1

## Beschreibung

**[0001]** Konventionelle Suchmaschinen arbeiten in der Regel auf dem Prinzip der Volltextindexierung. Bei der Volltextindexierung wird pro Dokument die Häufigkeitsverteilung von Begriffen des Dokuments oder eines Teils des Dokuments in einem invertierten Index erfasst. Dieser Index wird benutzt, um zum Anfragezeitpunkt die Dokumente zu bestimmen, in denen die gesuchten Begriffe auftreten. Des weiteren wird an Hand einer systemspezifischen Relevanzbewertungsfunktion für jedes Dokument ein Relevanzwert ermittelt. Auf der Basis der Relevanzwerte werden die Ergebnisdokumente anschließend sortiert ausgegeben.

**[0002]** Wesentlich hierbei ist die Tatsache, dass zur Bewertung nur die Begriffe herangezogen werden, die auch im Dokument auftreten.

**[0003]** Bei der Relevanzwertberechnung können bestimmte Elemente des Dokuments stärker gewichtet werden als der normale Textinhalt. Hierzu zählen:

- Meta-Informationen, insbesondere werden Inhaltsbeschreibende Stichworte ausgewertet
- Titel und Überschriften
  Die ersten Zeilen eines Dokuments
- Anzahl der Verweise auf das Dokument
- Ankertexte von Verweisen auf andere Dokumente
- Abstand zwischen Begriffen
- Phrasen

**[0004]** Die Ermittlung des Relevanzwertes erfolgt auf der Basis der relativen Häufigkeiten der Begriffe mit Hilfe von Informations-theoretischen Methoden. Kurze Dokumente, in denen die gesuchten Begriffe häufig auftreten, werden als relevanter bzgl. der angefragten Begriffe bewertet als längere Dokumente oder Dokumente, in denen die gesuchten Begriffe seltener auftreten. Entsprechend der informationstheoretischen Betrachtungsweise werden seltene Begriffe – bezogen auf den gesamten Dokumentenbestand – stärker gewichtet als Begriffe, die im gesamten Dokumentenbestand häufiger auftreten.

**[0005]** Verbunden mit diesem Ansatz sind folgende Probleme:

- Reine Volltextindexierung wurde für kleine, kontrollierte Dokumentenmengen konzipiert, die nicht notwendigerweise als verknüpfter Hypertext ausgelegt sind. Eine Übernahme der Volltextindexierung für Hypertexte (wie z. B. das World-Wide-Web (WWW) oder Web-basierte Intranets) nutzt die in den – in Hypertexten verwendeten – Verweisen kodierte Information nicht aus.
- Es können lediglich Begriffe gesucht werden, die in den Dokumenten selber auftreten, bzw. für die mit Hilfe eines Thesaurus synonyme Begriffe bestimmt werden können, die in den Dokumenten auftreten.
- Das Vorkommen von Begriffen einer Anfrage in einem Dokument sagt in der Regel wenig bzgl. der Relevanz des Dokuments bezogen auf die Anfrage aus, da die Bedeutung der Begriffe nicht erfasst wird und damit auch keine Aussagen über die Bedeutung des gesamten Dokuments möglich sind. Um dieses Defizit auszugleichen, wurden Ansätze entwickelt, bei denen die Dokumentautoren die Bedeutung des Dokuments in Form von Meta-Beschreibungen annotieren und bei denen das Vorkommen der gesuchten Begriffe in den Meta-Beschreibungen stärker gewichtet wird und so zu einem höheren Relevanzwert führt.
- Der Dokumentenautor wird nicht alle möglichen Bedeutungen des Dokuments erfassen und somit wird das Dokument nur für die vom Dokumentenautor erfassten Bedeutungen als relevanter betrachtet werden als andere Dokumente.



Durch die höhere Gewichtung der Meta-Beschreibungen ist die Relevanzbewertung bei unkontrollierten Dokumentenmengen offen für Manipulationen – als Spamming bezeichnet –, da die Dokumentenautoren willkürliche Begriffe in den Meta-Beschreibungen verwenden können.

**[0006]** Ein bekanntes Verfahren zur Relevanzbewertung bei der Indexierung von Texten basiert auf dem Lycos System. Bei dieser Lösung, die einer der ersten kommerziellen Suchmaschinen des WWW zugrunde liegt, wurden neben einer eingeschränkten Volltextindexierung, die lediglich die hundert "wichtigsten" Begriffe des Dokuments indexierte, zwei neue Konzepte eingeführt. Erstens, wurden Begriffe die in speziell ausgezeichneten Dokumentteilen auftraten (wie z. B. Titel, Überschriften, den ersten 20 Zeilen des Dokuments) bei der Relevanzbewertung stärker gewichtet als bei ihrem Auftreten in anderen Bestandteilen des Dokuments. Zweitens, floss in die Relevanzbewertung eines Dokuments bzgl. der Suchanfrage zum erste Mal eine Information über die "Dokumentenumgebung" in Form der "Anzahl der externen Verweise auf das Dokument" – als Popularität bezeichnet – mit ein, so dass Ergebnisdokumente, auf die sehr oft von anderen Dokumenten aus verwiesen wird, als "relevanter" betrachtet werden als Dokumente, auf die seltener verwiesen wird (Mauldin 97).

**[0007]** Die "Anzahl der externen Verweise auf ein Dokument" kann als eine Form eines "citation index" betrachtet werden, mit dem zwar in einigen Fällen die Qualität des Suchergebnisses verbessert werden kann, welches aber nicht in allen Fällen funktioniert. So werden beispielsweise bei einer Suche mit Lycos nach den Begriffen "Deutsche Telekom" ältere Presseveröffentlichungen als "populärer" betrachtet als die Homepage der Deutschen Telekom, auf die mit großer Wahrscheinlichkeit weitaus öfter verwiesen werden dürfte. Insofern erscheint die veröffentlichte Aussage über die Berücksichtigung der Popularität als fragwürdig.

**[0008]** Darüber hinaus werden hierdurch Meta-Beschreibungen des Inhalts nur im Rahmen der Methoden der eingesetzten eingeschränkten Volltextindexierung berücksichtigt.

**[0009]** Bekannt ist weiterhin ein mit Rankdex bezeichnetes Verfahren. Mit Rankdex wurde eine erste experimentelle Implementierung (http://rankdex.gari.com/) einer neuen Relevanzbewertungsfunktion veröffentlicht, welche auf dem Prinzip des "Hyper Vektor Votings" (HVV) basiert (Li 98). Bei dieser Bewertungsmethode werden sowohl die Popularität als auch die "Texte – als Ankertexte bezeichnet –, die in externen Verweisen auf ein Dokument verwendet werden" berücksichtigt, so dass "Dokumente, auf die häufig mit den gesuchten Begriffen verwiesen wird" als relevanter betrachtet werden als "Dokumente, auf die seltener mit den gesuchten Begriffen verwiesen wird". Der Inhalt der Dokumente wird bei dieser Methode – bis auf die Ankertexte nicht berücksichtigt.

**[0010]** Diesem Verfahren liegt die Beobachtung zu Grunde, dass Dokumentautoren, die auf ein anderes Dokument verweisen, den Verweis in den meisten Fällen mit einer kurzen und sehr prägnanten Beschreibung versehen, die den Inhalt des Dokuments, auf das verwiesen wird, sehr gut beschreibt, so dass der verwendete Ankertext als Meta-Beschreibung betrachtet werden kann. Wird beispielsweise ein Verweis mit den Begriffen "Deutsche Telekom" versehen, so wird man durch den Verweis in den meisten Fällen auf die Homepage der Deutschen Telekom verwiesen werden.

**[0011]** Die Meta-Beschreibungen der Ankertexte werden in der Regel von einer Vielzahl von Autoren erzeugt, wobei diese durchaus auch alternative Begriffe in den Ankertexten verwenden werden. So ist es beispielsweise möglich, dass auf die "Homepage" der Deutschen Telekom im WWW auch mit den Ankertexten "Homepage der Deutschen Telekom", "Deutsche Telekom AG", "Telekom", "German Telekom" etc. verwiesen wird. All diese Ankertexte können als alternative Meta-Beschreibungen betrachtet werden.

**[0012]** Die Gefahr des Spammings ist zwar auch bei diesem Ansatz gegeben, da prinzipiell ein Dokumentautor durch die gezielte Verwendung von bestimmten Ankertexten die Relevanzbewertungsfunktion manipulieren kann. Dennoch ist der Einfluss dieser Form des Spammings auf die Relevanzbewertungsfunktion jedoch vergleichsweise gering, da sie durch die Anzahl und Art der Ankertexte, die von anderen Autoren verwendet werden, nivelliert wird.

**[0013]** Mit dieser Form der Relevanzbewertung ist es darüber hinaus möglich, auch Dokumente zu finden, in denen die Suchbegriffe selber nicht auftreten, die aber mit den Suchbegriffen beschrieben werden können. Des Weiteren können auch Dokumente in anderen Sprachen gefunden werden, bzw. Dateien mit nicht-textuellem Inhalt, wie z. B. Bild-, Audio-, Video-, oder Archivdateien oder ausführbare Programme.

**[0014]** Der Rankdex Ansatz ist jedoch dadurch beschränkt, dass er den eigentlichen Inhalt der Dokumente nicht berücksichtigt.

**[0015]** Bei Rankdex handelt es sich um eine experimentelle Implementierung einer Suchmaschine, die auf HVV basiert. Zu Testzwecken wurden bei diesem Experiment 19975.3 Millionen Internetseiten gesammelt und ein Index von rund 100 MB aufgebaut. Durch einen Vergleich mit anderen Suchmaschinen konnte nicht nur die Qualität der Ergebnisse unter Beweis gestellt werden, es konnten ebenfalls die Vorteile und die der bereits oben beschriebene Nachteil identifiziert werden. Rankdex konnte bisher nicht inspiziert oder getestet werden, da die publizierte URL http://rankdex.gari.com/ bisher nicht zugreifbar war.

**[0016]** Mit dem Ansatz von Google (Bryn & Page, 98) wurde eine Methode vorgestellt, mit der die Nachteile reiner Volltextindexierung, der alleinigen Beurteilung der Popularität und der Ankertexte behoben werden.

**[0017]** Der mit Google vorgestellte Ansatz beruht darauf, dass alle zu verarbeitenden Dokumente aus dem WWW geladen und lokal gespeichert werden. Aus diesen Dokumenten wird die topologische Verweisstruktur extrahiert und ebenfalls gespeichert. Mit einer Bewertungsfunktion wird der "sogenannte PageRank" mit Hilfe eines in mehreren Durchläufen konvergierenden, iterativen Algorithmus berechnet. Der PageRank eines Dokuments errechnet sich aus den PageRanks "aller Dokumente, die auf das Dokument verweisen" und betrachtet lediglich die topologische Verweisstruktur und nicht den Inhalt der Dokumente. Da eine Rückwärtsverfolgung von Verweisen im WWW nicht möglich ist, kommt dieser Ansatz nicht umhin, alle Dokumente – resp. einen Großteil – zunächst zu laden und die topologische Verweisstruktur lokal zu speichern, bevor mit der Berechnung des PageRanks begonnen werden kann.

**[0018]** Bedingt durch die lokale Speicherung der Dokumente und der topologischen Verweisstruktur wird viel Speicherplatz benötigt.

**[0019]** Die Berechnung des PageRanks erfolgt dann selber in einem Stück, so dass der verwendete Algorithmus als "konpilierend" bezeichnet werden kann. (Bryn & Page 98) schreiben "a PageRank of 26 million web pages can be computed in a few hours". Zusammen mit einem anderen Prozess – als Sorter bezeichnet –, der rund 24 Stunden für die Sortierung dieser Datenmenge benötigt, benötigt der Aktualisierungsprozess von 26 Mio. Dokumente des Indexes demnach weit mehr als 24 Stunden. Wie dies zu der zuletzt geschätzten Indexgröße von rund 190 Mio. Dokumenten skaliert, und ob dies weiter optimiert wurde, ist unbekannt.

**[0020]** Zwar terminiert die Berechnung des PageRanks bei den Dokumenten, auf die von keinem anderen Dokument aus verwiesen wird, so dass deren PageRank prinzipiell als konstant betrachtet werden könnte. Das garantiert aber nicht, dass nicht irgendwann doch auf die Dokumente verwiesen wird, so dass die Berechnung des PageRanks bei einer Aktualisierung auch für diese Dokumente immer von Neuem erfolgen muss.

**[0021]** Bedingt durch den kompilierenden Ansatz bei der PageRank Berechnung kann eine Aktualisierung des Indexes nur in zeitlich größeren Abständen erfolgen.

**[0022]** In die eigentliche Berechnung des Relevanzwerts der Suchergebnisse fließen neben dem PageRank und den Standardmaßen des Information Retrievals weitere Informationen ein, wie z. B. das Vorkommen der Suchbegriffe im Titel, in Ankertexten, URLs oder speziell ausgezeichneten Textteilen und – bei Mehrwortanfragen – die Nähe zwischen den Vorkommen der einzelnen Begriffe. Wie diese Informationen miteinander verknüpft werden, ist nicht bekannt.

**[0023]** Bei Google handelt es sich um eine Internetsuchmaschine, die aus einem Projekt der Stanford University hervorging, welches 1998 in der Gründung der Firma Google, Inc. mündete. Aus der Zeit vor der Firmengründung sind detailliertere und publizierte Informationen über Google bekannt.

**[0024]** Bei Google werden wie bei Rankdex Ankertexte gesondert bewertet. Hierbei liegt der Unterschied der Verfahren, neben der gesonderten Bewertung anderer Textkomponenten, in der Bewertungsfunktion. Zwar wurde für Google diese Bewertungsfunktion nicht im Detail veröffentlicht, dennoch ist bekannt, dass sie neben dem Dokumentinhalt auch die Positionen der gesuchten Begriffe im Dokument, Formatierungsinformationen, Ankertexte und den PageRank des Dokuments miteinander kombiniert.

**[0025]** Der PageRank eines Dokuments ist ein globaler Wert, der unabhängig vom Inhalt allein aus der topologischen Struktur des WWWs bestimmt wird und als "Zitierungsgrad" interpretiert werden kann. Vereinfacht gesprochen erhalten Dokumente, auf die von "wichtigen" Dokumenten verwiesen wird, einen höheren PageRank als Dokumente, auf die von "unwichtigen" Dokumenten verwiesen wird. Je öfter auf ein Dokument verwiesen wird, desto "wichtiger" wird es eingestuft.

**[0026]** Der PageRank kann allein aus der topologischen Struktur, der Anzahl der Verweise und dem PageRank anderer Dokumente bestimmt werden. Zur Berechnung des PageRank eines Dokuments wird der PageRank aller Dokumente verwendet, die auf das Dokument verweisen. Zur korrekten Berechnung des rekursiv definierten PageRanks eines Dokuments muss somit der PageRank der auf sie verweisenden Dokumente bekannt sein.

**[0027]** Hieraus ergibt sich konsequenterweise der Schluss, dass bei einer Änderung des PageRanks eines Dokuments nicht nur dessen PageRank aktualisiert werden muss, sondern auch der PageRank aller von diesem Dokument aus erreichbaren Dokumente. Im schlimmsten Fall muss bei der Änderung eines Dokuments der PageRank aller Dokumente des Index neu berechnet werden.

**[0028]** Für Google wurde nicht beschrieben, wie die Bewertungsfunktion die einzelnen bewerteten Informationen kombiniert. Insofern ist auch unklar, wie Informationen aus dem Ankertexten mit dem PageRank kombiniert werden. Den Publikationen über Google kann entnommen werden,

dass eine Änderung von Dokumenten zwar permanent in den Index aufgenommen wird, die Berechnung des Page-Ranks und die Sortierung des Index jedoch in einer Stapelverarbeitung (Batch-Lauf) erfolgt, die allein für die parallele Sortierung von 24 Mio. Dokumenten auf vier Rechnern rund 24 Stunden benötigt. Hieraus ergibt sich die Folgerung, dass ein Index-Update als Stapelverarbeitung durchgeführt wird, und somit Indexaktualisierungen nur in zeitlich größeren Abständen erfolgen.

**[0029]** Die Erfindung ist auf ein Relevanzbewertungsverfahren ausgerichtet, dass eine bessere und aktuellere Indexierung von Hypertext-Dokumenten ermöglicht.

**[0030]** Grundlage des erfindungsgemäßen Verfahrens ist eine Suchmaschine, die nachfolgend mit "TeleFinder" bezeichnet wird.

**[0031]** Die Suchmaschine TeleFinder besteht so wie die meisten bekannten Suchmaschinen, im Wesentlichen aus zwei Komponenten, einem Robotersystem inklusive Datenbank und einem Indexserver inklusive Benutzeroberfläche.

**[0032]** Das Robotersystem lädt ausgehend von Startadressen Dokumente, durchsucht sie auf bisher unbekannte Dokumentadressen und übergibt die Dokumente dem Indexserver. Ausgehend von den neuen, unbekannten Adressen werden die korrespondierenden Dokumente geladen und der Zyklus erneut durchlaufen, bis alle erreichbaren Dokumente verarbeitet wurden.

**[0033]** Der Indexserver analysiert den Inhalt der Dokumente und baut einen invertierten Index auf, welcher für die Anfragebearbeitung benutzt wird. Wie bei jeder anderen Suchmaschine auch, wird die Qualität der Suchergebnisse durch die Inhalte der Dokumente, die Berücksichtigung ausgewählter Strukturelemente und insbesondere auch durch die verwendete Berechnungsfunktion bestimmt.

**[0034]** Das erfindungsgemäße Relevanzbewertungsverfahren für den Indexierungsvorgang des TeleFinders basiert auf der Grundidee die aus dem "Hyper Vector Voting" (HVV) bekannte Verfahrensweise der Ermittlung des Relevanzwertes eines Dokumentes anhand der Ankertexte von Verweisen, die auf das Dokument verweisen, mit der aus der konventionellen Volltextindexierung bekannten Verfahrensweise, die auf der Indexierung von Suchbegriffen aus dem eigentlichen Dokument basiert, zu kombinieren. Das erfindungsgemäße Verfahren bewirkt eine neue Qualität bei der Suche nach relevanten Dokumenten, da es die positiven Eigenschaften des Hyper Vector Voting Verfahrens mit den positiven Eigenschaften des Verfahrens der konventionellen Volltextindexierung in einem neuen Verfahren vereinigt.

**[0035]** Gegenüber herkömmlicher Volltextindexierung fließen durch die besondere Berücksichtigung und Gewichtung von Ankertexten (der Texte, mit denen die Verweise auf ein Dokument versehen werden) in die Gesamtbewertung auch Inhaltsbeschreibungen ein, die von anderen Dokumentenautoren erstellt wurden. Die Ankertexte, die meist sehr prägnant und präzise den Inhalt des referenzierten Dokuments beschreiben, bilden so eine Form von Meta-Beschreibung, die bei der Bewertung berücksichtigt wird.

**[0036]** Zur Relevanzbewertung benutzt der TeleFinder ein Relevanzbewertungsverfahren, welches durch Gewichtung unterschiedlicher Anteile der Funktion parametrisiert werden kann. Unterschiedlich gewichtet werden kann so der Einfluß der folgenden Dokumentenbestandteile auf den Gesamtrelevanzwert:

- Titel
- Überschriften unterschiedlicher Gliederungsebenen
- Phrasen
- Phrasen in Ankertexten
- Texte in Verweisen auf das Dokument
- Texte in Verweisen innerhalb des Dokuments
- Dokumentadressen

**[0037]** Durch unterschiedliche Gewichtung dieser Elemente ist die Relevanzbewertungsfunktion selber konfigurierbar.

**[0038]** Das erfindungsgemäße Relevanzbewertungsverfahren läuft in drei Phasen ab. Die drei Phasen müssen dabei nicht notwendigerweise sequentiell ablaufen.

**[0039]** In der ersten Phase, die mit Aufbauphase bezeichnet wird, liefert das Robotersystem Hypertext-Dokumente an den Indexserver. Der Indexserver analysiert den Inhalt der Dokumente nach drei unterschiedlichen Gesichtspunkten:

1. Werden Verweise in dem Dokument identifiziert, so wird für jede aus diesen Verweisen bestimmbare Adresse ein neuer Dokumenteneintrag im Index angelegt, sofern ein solcher noch nicht existiert. Ansonsten wird der Dokumenteneintrag entsprechend aktualisiert. Für die in den Verweisen verwendeten Begriffe der Ankertexte werden neue Termeinträge im Index angelegt, sofern diese noch nicht existieren. Ansonsten werden die entsprechenden Termeinträge aktualisiert. Für jeden Begriff des Ankertextes wird entsprechend einer Gewichtung ein partieller Relevanzwert vorausberechnet.
2. Werden speziell markierte Textinhalte (z. B durch die HTML Auszeichnungen Titel, H1, H2 oder H3 markiert) in dem Dokument identifiziert, wird für jeden Begriff, der in diesen markierten Textinhalten verwendet wird, ein neuer Termeintrag im Index angelegt, sofern dieser noch nicht existiert. Ansonsten werden die entsprechenden Termeinträge aktualisiert. Für jeden identifizierten Begriff wird entsprechend der Gewichtung der Markierung ein partieller Relevanzwert vorausberechnet.
3. Für jeden anderen nicht-markierten Textinhalt wird ein neuer Termeintrag im Index angelegt, sofern dieser noch nicht existiert. Ansonsten wird der entsprechende Termeintrag aktualisiert. Für jeden dieser Begriffe wird ein partieller Relevanzwert vorausberechnet.

**[0040]** In der zweiten Phase, die als Aktualisierungsphase bezeichnet wird, werden Dokumente deren Inhalte sich seit dem letzten Besuch verändert haben, zunächst aus dem Dokumenteindex entfernt. Die betreffenden Termeinträge werden aktualisiert. Sofern das veränderte Dokument weiterhin verfügbar ist, wird es entsprechend den Arbeitsschritten der Aufbauphase in den Index eingefügt.

**[0041]** Diese Verfahrensweise hat u. a. den Vorteil, dass ein Dokument – solange es sich nicht verändert hat – nur einmal über das Netz von einem anderen Server geladen werden muss, und dass es nicht lokal gespeichert werden muss. Darüber hinaus ermöglicht diese Verfahrensweise auch Verweise auf Dokumente in den Index aufzunehmen, deren zugehörige Dokumente selber noch nicht geladen und analysiert wurden.

**[0042]** Darüber hinaus wird durch die partielle Vorausberechnung der Relevanzwerte die Bestimmung des Relevanzwerts zum Anfragezeitpunkt minimiert. Mit unterschiedlichen Gewichtungswerten für Ankertexte in und auf Dokumente, für Phrasen und für unterschiedliche Textmarkierungen, ist die Relevanzbewertungsfunktion parametrisierbar und somit flexibel konfigurierbar.

**[0043]** Die dritte Phase wird mit Anfragephase bezeichnet.

**[0044]** In der Anfragephase werden in Abhängigkeit vom

verwendeten Anfragetyp (einfache Anfrage, komplexe Anfrage, Boot'sche Anfrage oder Phrasenanfrage) aus dem Index die Dokumente ermittelt, die auf die Anfrage zutreffen. Für jedes gefundene Dokument wird der eigentliche Relevanzwert aus den vorausberechneten Relevanzwertanteilen, der zum Anfragezeitpunkt vorliegenden Anzahl an Verweisen auf das Dokument und der Gesamtanzahl der Dokumente im Index zum Relevanzwert des Dokuments verrechnet.

**[0045]** Im Gegensatz zu dem mit Google vorgestellten Ansatz handelt es sich bei der im TeleFinder realisierten Lösung um ein inkrementelles Verfahren, bei dem aktualisierte Dokumente direkt in den Index integriert werden und somit prinzipiell umgehend nach einem als "flushen" bezeichneten Speichern des Indexes zur Suche bereitgestellt werden. Im Vergleich zu dem Google Ansatz kann dadurch eine weitaus höhere Aktualität des Indexes garantiert werden. Durch die direkte inkrementelle Verarbeitung von neuen bzw. aktualisierten Dokumenten müssen keine lokalen Kopien der Dokumente gespeichert werden, so dass der benötigte Plattenspeicherplatz drastisch reduziert werden kann.

**[0046]** Gegenüber dem Rankdex Verfahren verhält sich TeleFinder wie eine konventionelle Volltextsuchmaschine, sofern die gesuchten Begriffe nicht in Ankertexten auftreten. Das liegt darin begründet, dass auch der Inhalt der Dokumente indexiert wird.

**[0047]** Zwar wird im erfindungsgemäßen Relevanzbewertungsverfahren wie auch im Lycos-Verfahren die Popularität der Ergebnisdokumente bewertet, jedoch geht die Bewertung nach der erfindungsgemäßen Lösung weiter als beim Lycos-Verfahren, da neben der reinen Volltextindexierung, der Berücksichtigung spezieller Dokumentenbestandteile und der Popularität, wie bei Rankdex und Google auch die Ankertexte berücksichtigt werden.

**[0048]** Die Relevanzbewertungsfunktion ist darüber hinaus parametrisiert, so dass die einzelnen bei der Bewertung berücksichtigten Bestandteile unterschiedlich gewichtet und die Bewertungsfunktion insgesamt beeinflusst werden kann

## Patentansprüche

1. Verfahren zur Relevanzbewertung bei der Indexierung von Hypertext-Dokumenten mittels Suchmaschine, bei dem Hypertext-Dokumente in der Indexierungskomponente der Suchmaschine ausgewertet werden, **dadurch gekennzeichnet**, dass es in eine Aufbauphase, eine Aktualisierungsphase und eine Anfragephase unterteilt ist,

dass in der Aufbauphase die Hypertext-Dokumente in der Indexierungskomponente gleichzeitig auf das Vorhandensein von Verweisen, speziell markierten und nichtmarkierten Textinhalten durchsucht werden, wobei

a) bei der Identifizierung von Verweisen, für jede aus diesen Verweisen bestimmbare Adresse ein neuer Dokumenteneintrag in der Indexierungskomponente angelegt bzw. ein bereits vorhandener Dokumenteneintrag aktualisiert wird, dass für die in den Verweisen verwendeten Begriffe der Ankertexte ebenfalls ein neuer Termeintrag in der Indexierungskomponente angelegt wird bzw. ein bereits vorhandener Termeintrag aktualisiert wird, und dass für jeden Begriff des Ankertextes ein partieller Relevanzwert vorausberechnet wird,

b) bei der Identifizierung von speziell markierten Textinhalten, für jede ermittelte Markierung ein neuer Termeintrag in der Indexierungskomponente angelegt bzw. ein bereits angelegter Termeintrag aktualisiert wird, dass für jeden markierten Begriff ein partieller Relevanzwert vorausberechnet wird, und

c) bei der Identifizierung von nicht-markierten Textinformationen in einem auszuwertenden Dokument ein neuer Termeintrag in der Indexierungskomponente angelegt bzw. ein bereits zu der Textinformation vorhandener Termeintrag aktualisiert wird, und dass für jeden Termeintrag ein partieller Relevanzwert vorausberechnet wird,

dass in der Aktualisierungsphase bereits erfasste und indexierte Dokumente, deren Inhalt sich geändert hat, automatisch aus dem Dokumentenindex gelöscht werden,

dass die Termeinträge zu diesen Dokumenten aktualisiert werden, und dass die geänderten Dokumente sofern sie weiterhin verfügbar sind, noch einmal entsprechend der Aufbauphase in der Indexierungskomponente erfasst werden, und

dass in der Anfragephase auf die Anfrage eines Nutzers in Abhängigkeit vom Abfragetyp, wie einfache Frage, komplexe Frage, Boot'sche Anfrage bzw. Phrasenanfrage aus der Indexierungskomponente Angaben zu relevanten Dokumenten ermittelt werden, wobei für jedes ermittelte Dokument der eigentliche Relevanzwert aus den vorausberechneten Relevanzwertanteilen, der zum Anfragezeitpunkt vorliegenden Anzahl an Verweisen auf das ermittelte Dokument und der Gesamtzahl der Dokumente in der Indexierungskomponente berechnet wird, und dass die entsprechend ihrer Relevanzbewertung geordneten Angaben zu den Dokumenten zusammen mit Zusatzinformationen an den Nutzer ausgegeben werden.

- Leerseite -